Spectrogram-based Detection of Signal Contour Tracks in Underwater Audio Recordings

Shyam Kumar Madhusudhana, Christine Erbe, Alexander Gavrilov
Centre for Marine Science & Technology
Curtin University
Perth, WA, Australia
shyam.madhusudhana@ieee.org

Abstract— **Passive acoustic monitoring has long been used to detect various signals of interest in underwater audio recordings. In particular, systems have been developed for the detection of several types of narrowband signals corresponding to specific marine mammal vocalizations. In this work we present a generic system, based on post-processing of spectrograms, for the automatic extraction of frequency contours of narrowband signals produced by any source – biological, anthropogenic or physical ambient.**

Keywords—spectrogram; narrowband signal; contour extraction; passive acoustic monitoring; algorithm; underwater recording

I. INTRODUCTION

Given the high computational capability of modern computers and the availability of efficient implementations of fast Fourier transform (FFT), the use of spectrograms for analysis of underwater audio signals is a very attractive choice. Several approaches are available for detecting dolphin whistles, narrowband whale calls, tonal components of vessel noise, etc. in underwater audio recordings. Each approach is usually targeted at identifying one or more specific types of sounds and they exhibit the claimed levels of performance only in specific recording scenarios. A recording scenario involves the specific marine environment, its hydro- and geoacoustic characteristics, its ambient soundscape, the time of year, the recording equipment configuration and mooring, etc. all of which can affect the performance of a signal detector.

Examples of underwater tonal signals include odontocete whistles, mysticete vocalizations, quasi-tonal low-frequency sounds from ice events (such as drifting icebergs) [1, 2], sonar probes and vessel noise. Other tonal signals of biological origin worth mentioning are some fish sounds. Independent calls originating in close proximity to recording equipment are visually distinguishable in spectrograms, while choruses (highly overlapping calls) appear to have energy spread across several contiguous frequency bins rendering automatic detection problematic with spectrogram-based contour extraction methods.

Some of the existing contour-tracking approaches based on post-processing of spectrograms involve the use of image-processing operations such as image-thresholding and edge-detection among others (Ex: [3, 4]). Image-processing, due to its response latency, is too slow for real-time analysis as in onsite operations, making contour tracking methods based on image-processing of spectrograms better suited for offsite analysis. The approaches described in [2-4] tackle the contour extraction problem in two phases where prominent spectral peaks in individual frames (time bins) of a spectrogram are identified in the first phase following which closely lying peaks from neighboring frames are identified to eventually trace the underlying tonal signals' contours. The specifics of the peak-picking and the subsequent contour-tracing sub-processes vary in those approaches.

The contour extraction system presented here is a generic contour extractor and not targeted for any specific type(s) of contour(s). However, it also follows the two-phase approach (as those in [5-7]). This is part of a broader two-stage detection-classification system where the first stage (which is the scope of this article) involves only the detection/extraction of tonal signals present in audio recordings and the second stage performs classification upon examining the characteristics of the extracted contours. In consideration of the broader goal, our approach to contour extraction places due emphasis on engineering the solution, with a view to achieving faster than real-time performance with negligible response latency, while minimizing variations in performance under different recording scenarios. While we strive to achieve a high rate of true detections (recall), keeping the rate of spurious detections (false positives) to a bare minimum is important.

In view of the above considerations, the extractor presented in this article tackles both the peak-picking and contour-tracing sub-processes differently as compared to [5-7]. The proposed peak selection sub-process is an altered version of that in [6]. While retaining the simplicity, the alterations enable tracking of not only high-intensity signals but also those signals with low signal-to-noise-ratios (SNR). In the contour-tracing process, considering additional features such as the variations in contour slopes and in energy content of the underlying signals simplifies tracking of concurrent contours and aids in the separation of overlapping and intersecting contours. Unlike some of the published works on the extraction of *time x frequency* contours in spectrograms (Ex: [3-7]), since the goal in the current approach is to simply trace as many tonal signals as there are in the recordings regardless of the sources that produced the underlying sounds, only non-tonal signals (broadband sounds and white noise) are considered as

978-0-933957-40-4 ©2013 MTS

This is a DRAFT. As such it may not be cited in other works. The citable Proceedings of the Conference will be published in IEEE Xplore shortly after the conclusion of the conference.

"noise". Hence, data-conditioning operations such as spectral-means-subtraction (as used in [6, 7]) are not employed, which may otherwise jeopardize the detection of certain types of tonal signals such as long-duration ship sounds and certain polar ice events.

II. Overview of the Contour Extraction System

Spectrograms computed with carefully selected parameters present a simple and elegant means to visually distinguish tonal signals amongst other sounds. It would seem that this nature of spectrograms could be easily exploited for the purposes of automatic detection of tonal signals. However, the uncertainty principle resulting in a trade-off between the time and frequency resolution dictates that a single set of spectrogram parameters cannot yield similar distinguishing capability across wide ranges in the frequency domain. Hence, it is beneficial to compute multiple spectrograms with different sets of parameters for different frequency ranges. For all frames in a computed spectrogram, the peak-picking and contour tracing operations are restricted to only those contiguous values on the frequency axis that lie within the chosen frequency range. We will use the term *segment* to refer to such a frequency-range restricted portion of a spectrogram.

The proposed algorithm is generic, in that the functionality or control-flow would remain the same for the processing of any arbitrary *segment* in consideration and only certain algorithm parameters would need to be altered for efficient tracking of contours in the chosen frequency range. With the available range of frequencies for any recording, one may either wish to extract tonal signal contours within only one or more frequency ranges of interest or wish to extract all tonal signals contained in the recording. In the latter scenario, it is up to the user to carefully select *segment* boundaries (with or without overlap). The choices of the boundaries are influenced by three considerations:

- The need to maximize the efficacy of a set of values chosen for the spectrogram parameters imposes limitations on the width of the *segment*.
- On the contrary, excessive segmentation leads to undesirable increase in overall processing time..
- For effective contour tracing, the possibility of contours crossing boundaries has to be minimized. This necessitates prior examination of tonal signals in long-term recordings and an appropriate literature review before boundaries are chosen.

The actual choice of *segment* boundaries is often dictated by the purpose of analysis, the recording scenario and the repertoire of sounds that may be available in the recordings. Consider, for example, audio recorded with a sampling frequency of 8 kHz at a site having little anthropogenic activity. Barring a few types of mysticete and seal vocalizations, a frequency of about 100 Hz seems to provide a natural boundary separating seismic & ice events' sounds from marine mammal tonal calls (see [8-10]). Once suitable boundaries are chosen, processing of audio recordings for the extraction of tonal signals follows the operations described in the below subsections. In order to ease understanding, we will use a running example to simplify the description of the various components of the system. We have considered a ~10s long recording containing foraging calls from multiple killer whales. The audio, recorded at a sampling rate of 44100 Hz, contains harmonics of a few overlapping calls. A *segment* (768 Hz – 11000 Hz) of the example is shown in Fig 1.

A. Signal Preparation

Since all operations will be limited to the range of frequencies within a *segment*, filtering out energy from frequencies outside of this range before computing short-time Fourier transforms of the signal helps eliminate the undesirable effects spectral leakage would have on subsequent computations. Based on the chosen frequency extents of the *segment*, one may employ a low-pass, a high-pass or a band-pass filter as appropriate.

When the upper extent of a *segment* is below the Nyquist frequency of the recording, one may choose to downsample the signal in order to achieve computation speedup. The number of Fourier transform points (NFFT) considered in computing a spectrogram can be reduced by roughly the same factor as that

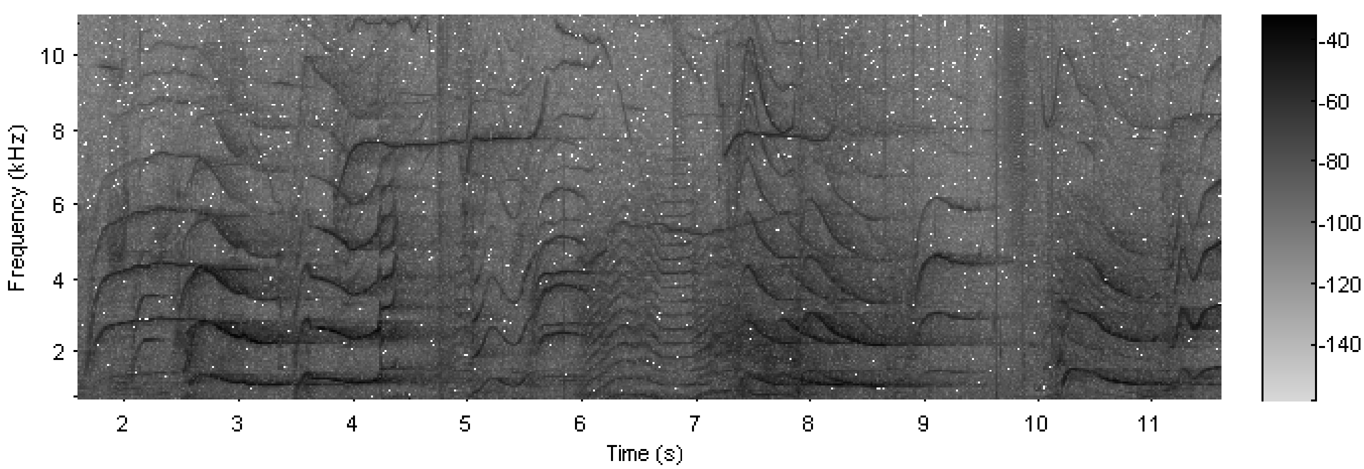

Fig 1. A *segment* (768 Hz - 11000 Hz) used in describing the contour extraction system. The spectrogram was computed using hamming windowing of ~35 ms long analysis windows with 50% overlap.

of downsampling without significant loss in frequency resolution of the ensuing spectrogram. When the parameters of the above filtering operation are so chosen as to suppress aliasing effects, the subsequent downsampling operation would not require explicit anti-aliasing filtering. In such a case, downsampling with an integer factor exhibits a constant-time performance.

For the example considered, the audio signal was bandpass filtered to suppress frequencies outside of the *segment* extents and then downsampled by a factor of 2 for a resulting sampling frequency of 22050 Hz. The spectrogram in Fig 1 was then formed from the log magnitude spectra of successive ~35 ms long (768 samples) Hamming-windowed frames with 50% overlap. The configuration provides 357 frequency bins per frame of the spectrogram with a frequency resolution (Δf) of 28.7 Hz and a time resolution (Δt) of 17.4 ms.

B. Peak-picking

The peak-picking process involves identifying significant spectral peaks in each frame of the spectrogram thus computed. A spectral peak is a local maximum which is a set of consecutive points that represent a maximum within its immediate neighbourhood. Presence of noise in audio recordings produces a large number of local maxima and not all such peaks are significant in view of tracing tonal signals. The approach in [6] retains only the highest few of such peaks for further processing. In the presence of multiple tonal signals, that approach risks discarding significant peaks. We present a simple approach to eliminate both less significant peaks and also those that correspond to broadband signals. The latter is important, as we are only interested in tonal signals.

Although they are regional maxima in a power spectrum, very low lying peaks do not stand out in a spectrogram. Peaks corresponding to broadband signals and those that appear dwarfed in a much broader neighbourhood also do not appear to stand out in a spectrogram. We employ a dynamic approach in eliminating such peaks. First, a smoothed version of the power spectrum is obtained. Let it be denoted as p_smooth_i, where i is the index of the point in the spectrum. The number of points considered in the smoothing operation is the next higher integer of the square-root of the total number of points in the spectrum (357 in the example, leading to 19 points for smoothing operation). The absolute difference between p_smooth_i and the mean of the spectrum (p_mean) is determined. Let it be denoted as $p\_smooth\_mean\_diff_i$. Finally a threshold value is determined as

$$p\_threshold_i = \max\{\, p\_mean\ ,\ [p\_smooth_i + (40\%\ of\ p\_smooth\_mean\_diff_i)\,]\,\} \quad (1)$$

p_smooth gives a measure of the general trend of energy spread in the spectrum. The 40% factor was obtained empirically and the addition of *p_smooth_mean_diff* to *p_smooth* is a way of favouring peaks that are considerably higher than the spectrum mean. Only those peaks that are higher than $p\_threshold_i$ are retained.

For all the remaining peaks, their frequency and power magnitude values are refined by quadratic interpolation of the three consecutive points that define the peak. Echoes with very short delays may cause peaks that are very close in frequency ([7]). In order to avoid detection of such duplicates, we discard the lower among them. Since the goal is to enable the overall system to work with arbitrary segments, we cannot use a fixed range in deciding whether or not neighbouring peaks are close enough to be considered as possible echoes. Instead, we compute this minimum distance between peaks as a function of the *segment*'s bandwidth and express it as

$$F\_GAP = \Delta f \times \log(\text{midpoint of segment's bandwidth}) \quad (2)$$

The log-form is an attempt to cater to the constant bandwidth-to-center-frequency ration. For the example considered $F\_GAP = 249.22$.

Fig 2 illustrates peak-picking on a frame (at 5.6076 s) from the spectrogram of Fig 1. Notice around 3200-3400 Hz that the presence of some broadband energy shows smeared energy in Fig 1. Also notice that there are several other peaks beyond 8kHz in the spectrum that do not stand out in the spectrogram of Fig 1. The described approach has successfully ignored all such corresponding peaks as can be seen in Fig 2. Fig 3 shows the result of applying the above described peak-picking process to the example considered.

In an onsite Passive Acoustic Monitoring (PAM) scenario,

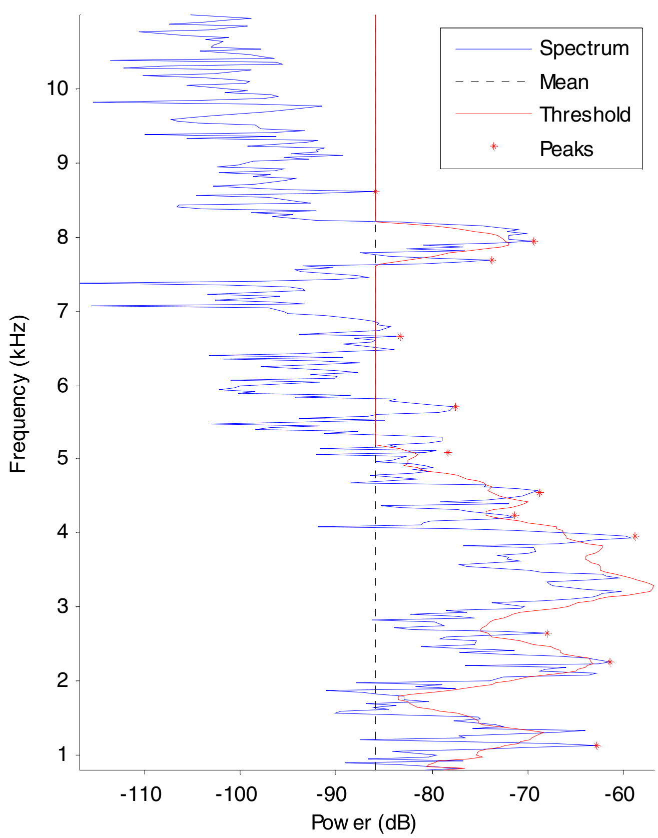

Fig 2. (Color online) Power spectrum (blue ilne) of an individual frame (at 5.6076 s in Fig 1), along with the corresponding dynamic threshold (red line) and the final set of quadratic-interpolated peaks.

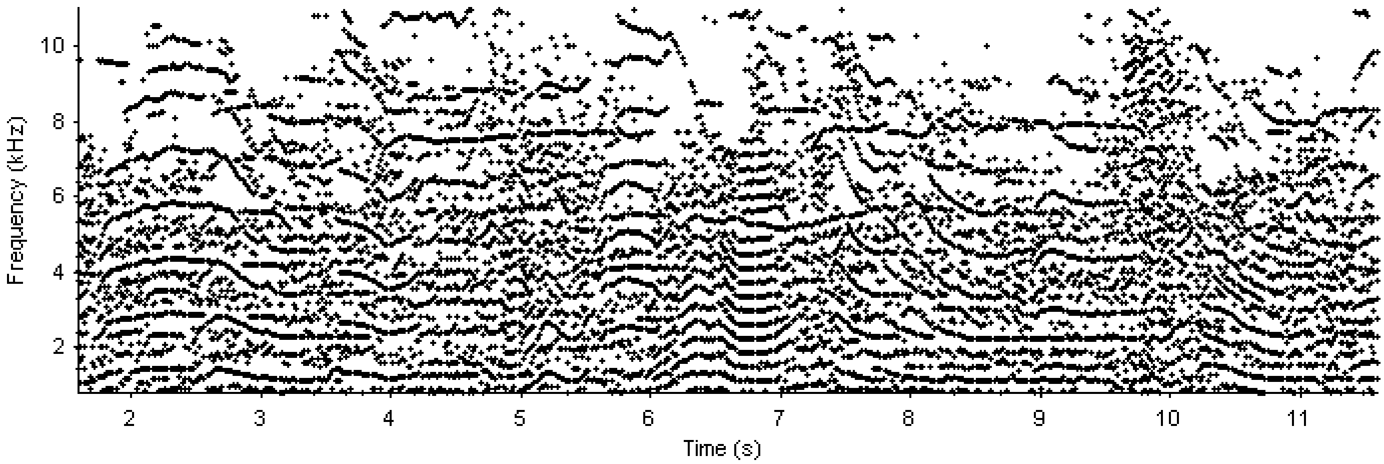

Fig 3. Result of peak-picking process performed on the spectrogram of Fig 1. Each retained peak is shown as a dot.

as chunks of audio samples become available, the above data preparation operations involving filtering, downsampling and spectrum computation and the subsequent peak-picking process can be performed in a pipelined fashion. For both onsite and offsite PAM scenarios, the peak-picking and the subsequent contour tracing operations are performed iteratively on successive frames of the spectrogram. Hereafter, the use of the term peak refers to ones that remain after the peak-picking process.

C. *Contour Tracing*

Peaks identified in successive frames that have a small separation on the frequency axis are associated. A thread of such successively connected pairs of peaks across multiple contiguous frames of a spectrogram constitutes an extracted frequency contour and this process of successively establishing associations is called tracing of the contour. A contour being traced is said to end when no further associations are possible. A contour that has not yet ended will be called an active contour. A peak in a given frame may be included in the traces of more than one contour. Some peaks, called orphan peaks, may not have any contours passing through it. Before we delve into the details of the contour tracing process, we would like to introduce a term called *min_contour_length*, expressed in time units (seconds). It sets the lower bound for the length of a traced contour. Following the rationale described in [7] for a corresponding term, we also set its value to 150 ms.

In associating peaks from neighbouring frames, the value of the upper bound on the separation between the peaks depends on whether or not the peak in the earlier frame is part of any active contours. If a peak in the earlier frame is an orphan, it will be connected to all those peaks in the succeeding frame that are within *F_GAP* distance on either side of the former. The *F_GAP* value is used again here for the same reasons as described earlier. All peaks in range are considered without any additional checks since our aim is to trace all tonal signals present. If any of these lead to false positives, they will be eliminated later.

When a peak in the earlier frame is not an orphan, then the active contour(s) passing through the peak has/have certain profile that can be exploited to minimize the number of candidate extensions. First, the slope of the active contour at its tip (two most recent points) determines the range in which candidate peaks from the succeeding frame may be considered. The previous lookup range of [-*F_GAP*, *F_GAP*] is now divided into three equal parts. The range covering the part in which the active contour's tip occurs plus its immediate neighbour constitutes the lookup range for the succeeding frame. Fig 4 illustrates how the lookup range is chosen. The motivation for this range-reduction scheme is to restrict further checks to only those peaks in a succeeding frame that do not produce a radical change in slope. Of the peaks that lie in the reduced range, those that result in a tip slope which is not within three standard deviations of the overall mean of the contour recorded so far are discarded. This eliminates peaks which seem like outliers in terms of overall slope since three times the standard deviation covers nearly 99.7% of the recorded values. Finally, we check the power profile at the head of the active contour. We define a contour head as (*min_contour_length* / 2) seconds of the most recent points of the contour. Standard error is computed for the distribution of power in the contour head (SE_h). Among the remaining

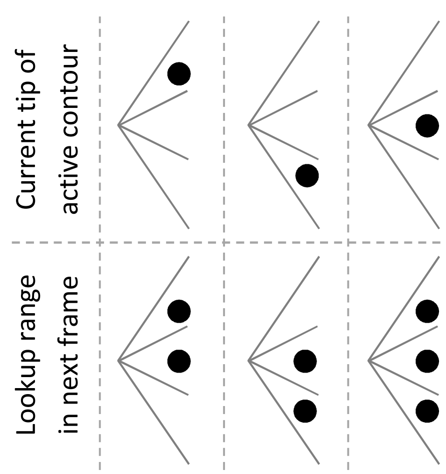

Fig 4. Graphic indicating how the new peak lookup range is chosen based on the tip of the active contour.

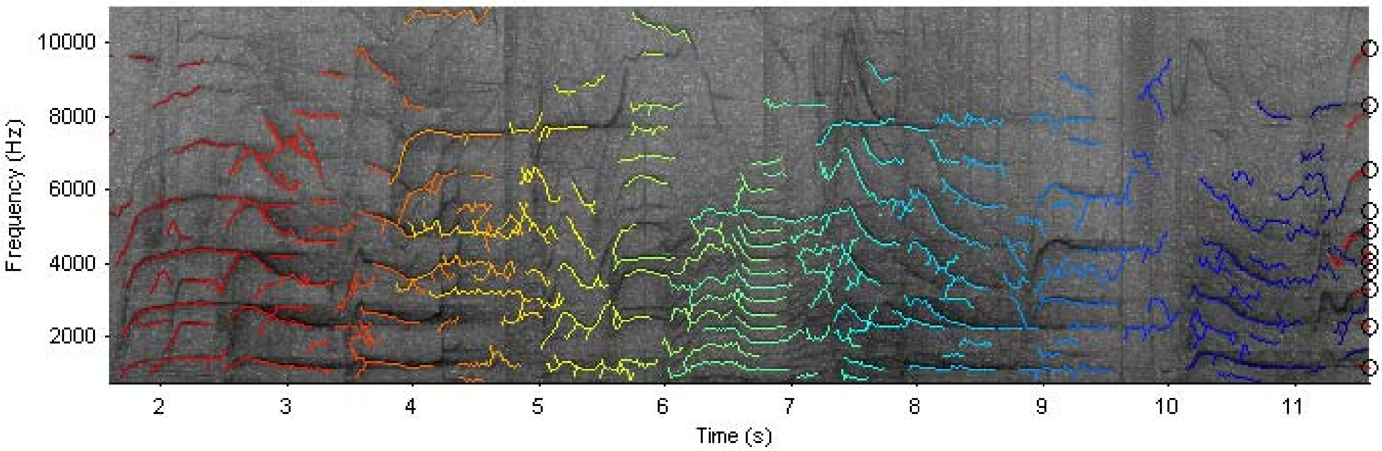

Fig 5. (Color online) The original spectrogram of the considered example along with overlays of the contours extracted as described. Individual contours are overlaid with different colors. The currently active contours are indicated by red dashed curves towards the right end of the spectrogram with a black circle indicating their tips.

candidate peaks, those whose inclusion in the contour would not yield a new standard error that is within 10% of SE_h are discarded.

When the tracing of a contour ends due to unavailability of candidate peaks for extension, those ended contours whose full duration is shorter than *min_contour_length* are rejected. Fig 5 shows the outcome of the overall process.

III. RESULTS

As we can see from Fig 5, the contour extractor described in this paper has shown good performance in extracting most of the contours in the considered example. However, one can notice that a couple of strong tonal signals have been missed – around 5 kHz at about 6s and also at 9s. Upon close examination, it was discovered that there were one or more very short discontinuities in tracing these calls. They were either from missing peaks or due to presence unsuitable peaks. The ensuing fragmentation caused the broken tracks to be shorter than *min_contour_length* and hence some of them were rejected.

IV. FUTURE WORK

Due to time limitation, a detailed performance analysis of the system could not be performed. The authors would continue to test the system to obtain vital performance statistics such as precision, recall and realtime factor, and make the results available in other ways.

Although it may not seem entirely appropriate given the fundamental differences in goals, the authors would like to perform comparative performance evaluation against some of the existing and publicly available automatic contour tracking methods such as those described in [7] & [5].

ACKNOWLEDGMENT

The authors would like to thank Chevron Australia who supported this work which is part of a larger project on characterizing marine soundscapes and monitoring anthropogenic noise.

REFERENCES

[1] B. Li and A. Gavrilov, "Localization of Antarctic ice breaking events by frequency dispersion of the signals received at a single hydroacoustic station in the Indian Ocean," *The Journal of the Acoustical Society of America,* vol. 123, pp. 2990-2990, 05/00/ 2008.

[2] J. Talandier, O. Hyvernaud, D. Reymond, and E. A. Okal, "Hydroacoustic signals generated by parked and drifting icebergs in the Southern Indian and Pacific Oceans," *Geophysical Journal International,* vol. 165, pp. 817-834, Jun 2006.

[3] S. Datta and C. Sturtivant, "Dolphin whistle classification for determining group identities," *Signal Processing,* vol. 82, pp. 251-258, 2002.

[4] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics,* vol. 32, pp. 39-47, Jun 2004 2004.

[5] X. C. Halkias and D. P. W. Ellis, "Call detection and extraction using Bayesian inference," *Applied Acoustics,* vol. 67, pp. 1164-1174, Nov-Dec 2006.

[6] S. K. Madhusudhana, E. M. Oleson, M. S. Soldevilla, M. A. Roch, and J. A. Hildebrand, "Frequency based algorithm for robust contour extraction of blue whale B and D calls," *Oceans 2008 - Mts/Ieee Kobe Techno-Ocean, Vols 1-3,* pp. 809-816, 2008.

[7] M. A. Roch, T. S. Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, and M. S. Soldevilla, "Automated extraction of odontocete whistle contours," *Journal of the Acoustical Society of America,* vol. 130, pp. 2212-2223, Oct 2011.

[8] C. Erbe, "The Acoustic Repertoire of Odontecetes as a Basis for Developing Automatic Detectors and Classifiers," DTIC Document2004.

[9] W. F. Perrin, B. Wursig, and J. Thewissen, *Encyclopedia of marine mammals*: Access Online via Elsevier, 2009.

[10] D. H. Thomson, *Marine mammals and noise*: Access Online via Elsevier, 1995.